Root Cause Analysis for PBI000000000057
CDF dCache prolonged outage incident 09/22/2009-09/23/2009

Report submitted 9/29/2009
(CD-doc-3485)

Summary of Incident

On Tuesday 09/22/2009 the off-line group of CDF Collaboration started scheduled downtime of CDF off-line computing systems. As a part of this downtime CDF dCache system was brought down in order to apply security kernel patches to the nodes running the system. After work was completed (afternoon of the same day) the dCache system was started up but remained down due to failure of one of its components - PnfsManager to reconnect to the system and failure of so called KDC multiplexers to start at the OS boot time. The latter prevented so called kerberised dcap doors, that are used to access data at CDF, from functioning.

These direct causes of the outage were identified and resolved by dCache developers after they were alerted on the next day (09/23/2009) that CDF dCache system was in trouble. The entire CDF dCache remained non operation until about 1:00 PM of 09/23/2009, a service interruption that seems unnecessarily long. Although this interruption was still within the boundaries of the contingency scenario of scheduled activities.

A first report of the problem with CDF dCache system was e-mail sent to dcache-amin@fnal.gov at 04:10 PM that indicated that something was wrong with the monitoring database. This e-mail did not receive immediate attention and later, already after business hours, an incident report with urgency level High was made using Service Desk requester console which is not a proper procedure for reporting a service disruption to a system on 24x7 support, because e-mail and web generated requests are handled only on 8x5 basis by the Service Desk. The correct procedure would be to alert a person at CDF who is authorized to request a page for dCache system. This person must have made a phone call to Service Desk with specific request to page SSA.

At 09:16 AM 09/23/09, the Service Desk routed the ticket to FEF (a decision based on keyword “CDF”). FEF promptly replied indicating that ticket needs to be re-routed to SSA. SSA received notification at 10:37 AM. At 11:26 AM SSA, following existing procedure of communicating issues they cannot resolve to dCache developers, cut a bugzilla ticket and replied to Service Desk that dCache developers were looking into the issue.

Between 11:26 AM and 1:00 PM the direct causes of CDF dCache outage were identified and addressed by restarting PnfsManager and starting KDC multiplexers on dcap door nodes.

List of related incident tickets:

- INC000000011514.

Background Concepts

The following points are useful in understanding the nature and impact of this problem.

System Overview

CDF dCache system consists of distributed services, written in Java. It exclusively provides location independent, mostly read-only access to 5.6PiBs of the experimental data to off-line systems (user analysis and production). dCache is developed by DESY/Fermilab/NDGF Collaboration. CDF is running version 1.7.0 of the product. Current “official” production version is 1.9.2-4. The essential components of dCache:

- Pools. Many pools run on so called pool nodes. A pool manages collection of data files located on a single disk partition, delivers file to user and stages it from tape as needed.
- PnfsManager. A single PnfsManager runs on pnfs node (cdfensrv1n). Provides namespace service to dCache system.
- PoolManager. A single PoolManager runs on a head node. Manages collections of pools.
- Doors. Many doors provide end user access to the system. Doors perform authentication and authorization functions as well as implement data access protocols (kerberised dcap, ftp, gridftp etc.). CDF uses mostly kerberised dcap protocol for data access. 30 kerberised dcap door run on 3 so called admin nodes.

A KDC multiplexer which is not a dCache component but a separate service that allows Java programs to spread their kerberos authentication requests across a suite of KDCs to lift the limitation of exactly one KDC imposted by the Java Kerberos API. Availability of this service is crucial b/c the access to data in CDF dCache is available almost exclusively via kerberised dcap protocol.

A client connects to the system via kerberised dcap door to retrieve/open a file; it gets authenticated using kerberos certificate (here the door uses KDC multiplexer). After that, the file metadata is queried from the PnfsManager to perform client authorization and to retrieve file unique identifier. This identifier is used to query PoolManager for a pool that contains the file and then the data is streamed from that pool directly to the client using dcap protocol. If file is not present in any pool, the client waits until the data is staged to available pool from tape.

The whole dCache system is normally stopped by executing a single script `cold-stop` by root user on monitoring node (cdfdcam). Start of the system is performed by a similar script called `cold-sart`. Both scripts are available in `~enstore/dcache-code/dcache-fermi-config/cdfen/` directory. The purpose of these scripts is to guarantee correct order of execution of start/stop procedures of various services the dCache relies on. Start/stop of KDC multiplexers is performed by these scripts. The PnfsManager is the only service not covered by these scripts b/c it runs on separate hardware managed by different group – SSA.

Dcache services running on a given node can be started stopped by executing:

```
/etc/init.d/dcache-core start|stop
/etc/init.f/dcache-pool start|stop
```

Both scripts are compatible with `chkconfig` utility but normally are disabled b/c distributed dCache requires correct order of startups of ancillary services.

Support Model

The system overall is on 24x7 support agreement. Pools are on 8x5 support. Dcache developers are responsible for installation of dCache system and ancillary services.

CDF DH experts perform stop/start procedures and modify system setup according to their needs consulting with dCache developers. For a number of years within dCache developers group there was a role of permanent CDF dCache liaison whose duties were to perform creation of CDF specific dCache core and pool RPMs, dCache installations, development and execution of stop/start procedures and troubleshooting. Just recently this role has been eliminated resulting in shifting of some of the tasks, previously performed by liaison, to CDF DH group.

The communication between CDF DH and dCache developers is done via SSA group which is responsible for system operations and provides 24x7 rotating shift service. Normally a mail from CDF DH would be sent to dcache-admin@fnal.gov and SSA primary will triage the problem, cutting if necessary bugzilla ticket with bug report for dCache developers. The same procedure is in place for incident tickets generated via Service Desk. A dCache developer primary checks bugzilla tickets and takes action or assigns ticket to the proper expert. If incident happens after business hours the SSA primary would either call dCache developer directly of via DMS department head.

The Service Desk handles web generated or e-mailed incident tickets according to their urgency levels during business hours on 8x5 basis. After hours a telephone call is required to initiate a page to the appropriate support group.

Timeline

Cast of characters:
Angela Bellavance (AB) CDF Data Handling head
John Hendry (JH) SSA primary
Stephan Lammel (SL) CDF off-line infrastructure head
Vijay Sekhri (VS) dCache developer primary
Rick Snider (RS) CDF Offline co-head
Rick StDennis (RSD) CDF Offline co-head
Timur Perelmutov (TP) dCache Project Leader
Eric Wicklund (EW) CDF Data Handling expert

09/22/2009

11:00:00 AM AB: Noticed lots of errors coming from the dCache system (may be related to downtime work)

03:22:00 PM SL: Work is complete. Request to bring dCache up

04:10:00 PM AB: Pools start up completed. There are errors. Mail to dcache-admin@fnal.gov with subject [Fwd: Monitoring database not being updated on Linux fcdfdcache14.fnal.gov ...].
Similar mails were received by dcache-auto@fnal.gov automatically

04:50:00 PM All pools are up. But still see errors. EW tried test job. Test job failed.

06:27:00 PM Created Service Desk ticket INC000000011514 with Urgency “High” that CDF dCache system is not working

07:54:00 PM AB: No response from either dcache-admin@fnal.gov or Service Desk. Logged in to head node fcdfdcache10 and restarted all dcache processes (/etc/rc.d/init.d/dcache-core restart). After that dCache cell monitoring page http://www-cdf.fnal.gov/samwww/prd/DHAtAGlance/dcache/cell_info.html cleaned up. PnfsManager stayed OFFLINE.

09/23/09

09:16:00 AM Service Desk assigned ticket INC000000011514 to FEF.

10:10:00 AM AB: checked status of Service Desk ticket.

10:17:00 AM SL: sends mail to dCache Project Leader (Timur Perelmutov) stating that CDF dCache is down since yesterday downtime, requesting help.

10:20:00 AM AB: send mail directly to dCache developer primary (VS) asking for help.

10:27:00 AM dCache developer primary reports failure to access fcdfdcache14, requests help from SSA

10:37:00 AM JH: SSA received Service Desk e-mail notification of incident ticket INC000000011514.

11:11:00 AM TP: forwards SL's ticket to dcache-admin stating that since bugzilla ticket was not opened he assumed that SSA is handling the issue.

11:26:00 AM JH: Created Bugzilla ticket 411 (http://www-ccf.fnal.gov/Bugzilla/show_bug.cgi?id=411).Priority P5, Severity “enhancement”.

11:31:00 AM SSA primary replied to Service Desk that issue has been passed on to dCache developers.

11:54:00 AM TP: successfully logged to CDF pnfs node (using hostname cdfensrv1n) and restarted PnfsManager.

12:20:00 PM AB: Get notice that PnfsManager restarted. Tried to start dCache services on fcdfdata10 (head node) by running /etc/rc.d/init.d/dcache-core restart.

12:24:00 PM AB: Still errors in the log. Doors still down. Sent mail to dcache-admin sayin so. Got reply from TP that he is looking. At this point RS, RSD, AB have TP on the phone while he debugs issue.

12:25:00 PM Bugzilla ticket 411 closed. Reply sent to Service Desk to ticket INC000000011514.

12:40:00 PM TP: discovered that KDC multiplexer failed to start. Started KDC multiplexer on head nodes. Restarted all kerberised dcap doors. CDF dCache system restored to full functionality.

01:14:00 PM EW runs tests. Tests successful.

Supporting Plots

Figure 1 shows transfer time of a test file using kerberised dcap door. A discontinuity in the graph corresponds to the downtime on 09/22/2009. Just after the downtime transfers hang till about 1:00 PM

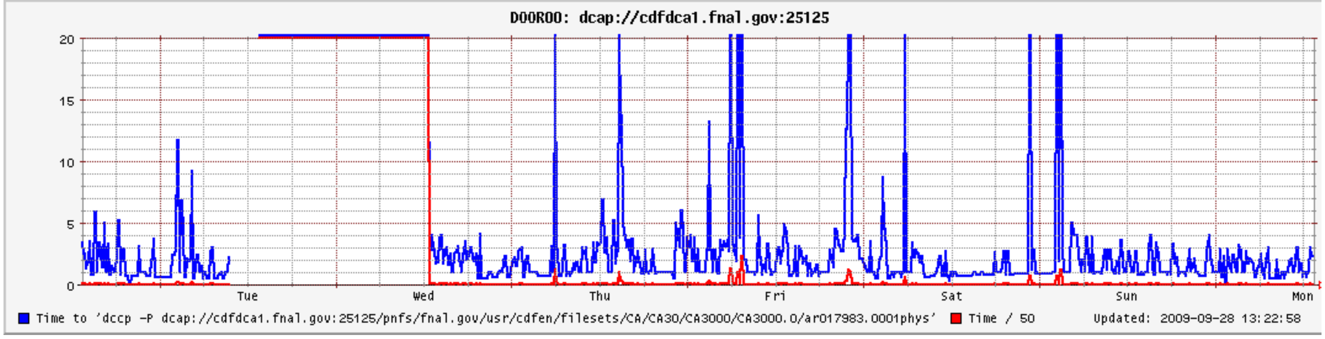

Figure 1: Dcap door monitoring plot shows test file transfer time in seconds as a function of day time. A discontinuity in the graph line corresponds to downtime on 09/22/2009. Just after the downtime transfers hang till about 1:00 PM on 09/23/2009.

on 09/23/2009. Plots for all 30 doors looked similar showing that CDF dCache was not operational during this time period.

Analysis

The problem analysis is represented as a fishbone diagram in Figure 2 where the reasons that caused problems in each category are specified.

People:

As CDF DH group took more responsibility for running dCache system after dCache developer liaison role have been discontinued there was either not sufficient amount of knowledge transferred from liaison to CDF DH or this knowledge was not well documented. Therefore it could be concluded that CDF DH experts did not receive sufficient training to perform their duties. In addition, staff rotation at CDF created a situation, when previously accumulated knowledge was lost or became obsolete. In turn, the dCache developer primary during the incident was a fresh transfer from other department and apparently did not have sufficient training to tackle the problem, especially as it turned out he could not even login to the system (during troubleshooting on the next day 09/23/2009).

Specifically, during root cause analysis investigation, it was discovered that `cold-stop` and `cold-start` scripts were not used. According to the logs and history data the dCache nodes were simply rebooted following kernel upgrades. Since `dcache-core` and `dcache-pool` turned out to be enabled in runlevel 3 on all CDF dCache nodes, which is a misconfiguration of system on the part of dCache developers who performed the installation, the individual components of the system attempted to start up automatically resulting in a system in inconsistent state. The system was attempted to be stopped one component at a time and then restarted. Excerpt from history on cdfdca:

```
 433  /bin/sh /opt/d-cache/jobs/billing -domain=billing-fcdfdcache10Domain stop
 434  /bin/sh /opt/d-cache/jobs/billing -domain=billing-fcdfdcache10Domain start
 435  /bin/sh /opt/d-cache/jobs/statistics stop
 436  /bin/sh /opt/d-cache/jobs/statistics start
 437  /bin/sh /opt/d-cache/jobs/httpd stop
 438  /bin/sh /opt/d-cache/jobs/httpd start
 439  /bin/sh /opt/d-cache/jobs/httpd stop
 440  /bin/sh /opt/d-cache/jobs/httpd start
 441  /bin/sh /opt/d-cache/jobs/dCache stop
 442  /bin/sh /opt/d-cache/jobs/dCache start
 443  /etc/rc.d/init.d/dcache-core restart
```

indicating that after startup the individual components were stopped (b/c they must have been running after reboot due to dCache components start up scripts enabled in runlevel), then started and then all services on the node restarted.

KDC multiplexers were not part of automatic boot up procedure and therefore were not running after nodes were rebooted. Since `cold-start` was not run, the KDC multiplexers did not start. Nor were they started by hand.

Dcache developers and SSA group were not sufficiently aware of ongoing work on CDF dCache system. If they were, they might have reacted more promptly to the first mail that was sent to dcache-admin@fnal.gov on 09/22 indicating that monitoring database was in trouble. Here it should be noted that CDF off-line infrastructure head did alert DMS leaders of forthcoming CDF dCache downtime.

Technology:

CDF DH group brought the system down as planned according to their documented procedures. During downtime, the PnfsManager which runs on a separate pnfs node cdfensrv1n has not been stopped (or rather was not requested to be stopped since CDF DH do not have control over this component). Since the rest of dCache was unavailable during downtime, the PnfsManager lost contact with service locator (head node running dCache domain) and became disabled. On subsequent dCache startup the PnfsManager did not reconnect, and eventually was declared OFFLINE on the dCache cell services page (http://cdfdca.fnal.gov:2288/cellInfo). PnfsManager does not reconnect in version 1.7.0 whereas in newer dCache versions it will try to reconnect till infinity. Since CDF uses 1.7.0 the PnfsManager had to be restarted with the rest of the dCache system. Apparently this was not known to CDF DH group.

KDC multiplexers were assumed to come up after system reboot. But it was discovered that there were no runlevel scripts installed on the CDF dcap door nodes that would start KDC multiplexers. During initial stage of root cause analysis it was considered as a consequence of incomplete installation, but as was later determined was caused by not using `cold-start` procedure to start dCache.

Conversely, `dcache-core` and `dcache-pool` should not be enabled in any runlevels to prevent system from starting prematurely. The fact that these services were on in runlevel 3 indicate misconfiguration of CDF dCache installation performed by dCache developers.

Processes:

An improper procedure for reporting incidents on the systems with 24x7 support during Service Desk off-hours was chosen by CDF DH expert. Instead of arranging a phone call by a CDF person authorized to request page by Service Desk to SSA primary a web request has been made at 6:27 PM. This resulted in a major delay of about 15 hours before the Service Desk handled the ticket on the next day.

Service Desk started to work on the ticket at 9:16AM following their established procedure for handling high priority tickets (this was a day when there were many of these). Service Desk routed ticket to the wrong group (FEF) because they did not know that request with "dCache" keyword needs to be routed to SSA. As FEF replied to the ticket suggesting to assign it to SSA, it took another hour for Service Desk to rout it to SSA because FEF did not re-assign it back to CSI/Setvice desk group, so that the ticket would have gotten into higher priority queue ("assigned to Service Desk tickets queue") but remained a "working ticket" going in the end of appropriate queue. This delay might have not have occurred at all had FEF re-assigned the ticket directly to SSA. They could have done it since they knew what is the correct group for handling dCache issues.

It took about 1 hour for SSA primary to create bugzilla ticket, which is required, according to existing procedure, for dCache developers to start working on it. There is a caveat here. Since by this time the CDF DH expert communicated directly with dCache developer primary, and CDF offline leaders communicated directly with DMS department head and dCache project leader the issue of non-functioning dCache already propagated to dCache developers and SSA primary was busy with answering questions from both CDF DH and dCache developer primary who needed help with logging to the system.

After bugzilla ticket was cut it took dCache developers some time to figure out how to login to the system before they could effectively address the problems.

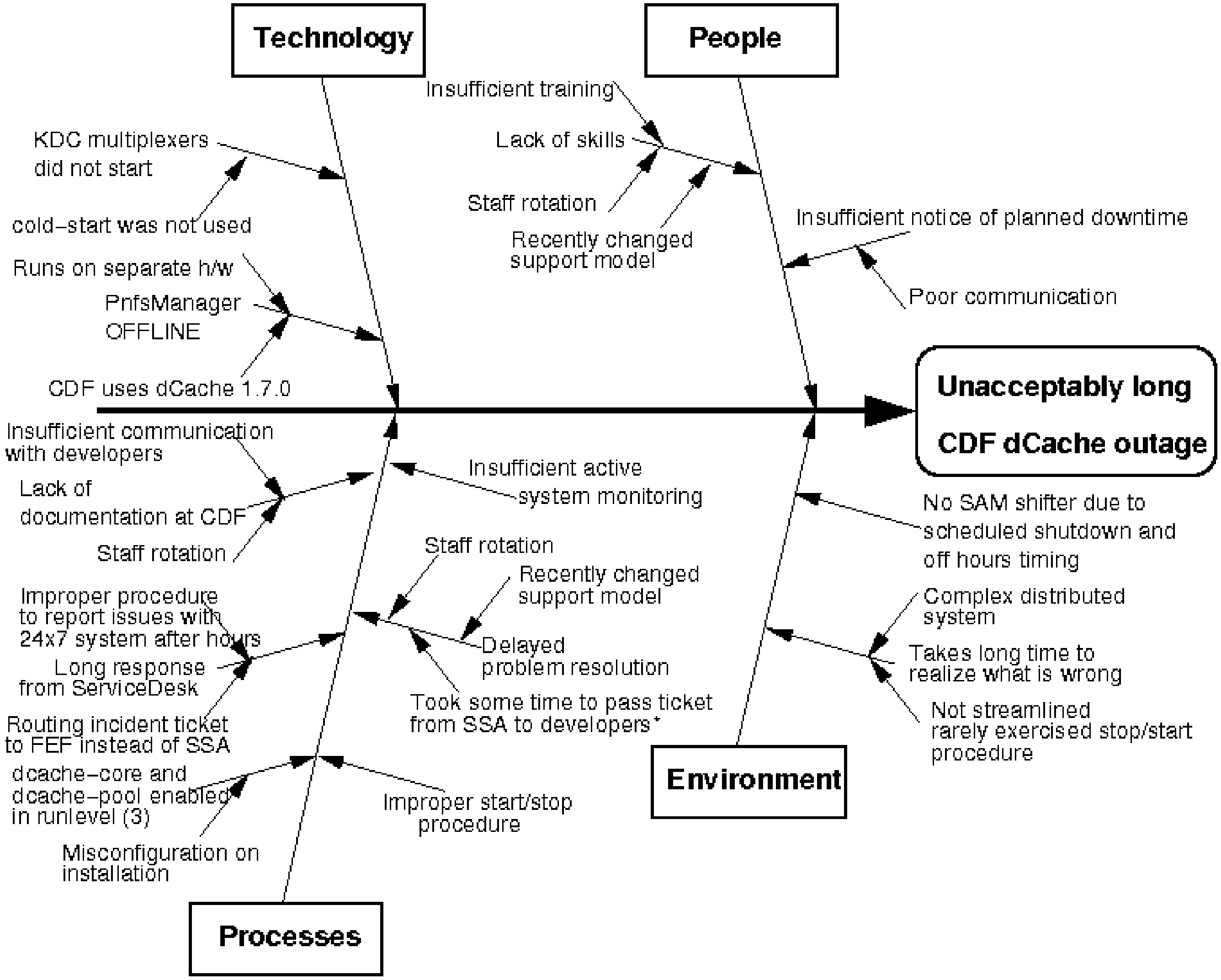
Technology
People
Insufficient training
Lack of skills
Staff rotation
Recently changed support model
KDC multiplexers did not start
cold-start was not used
Runs on separate h/w
PnfsManager OFFLINE
CDF uses dCache 1.7.0
Insufficient notice of planned downtime
Poor communication
Unacceptably long CDF dCache outage
Insufficient communication with developers
Lack of documentation at CDF
Staff rotation
Insufficient active system monitoring
No SAM shifter due to scheduled shutdown and off hours timing
Staff rotation
Recently changed support model
Delayed problem resolution
Took some time to pass ticket from SSA to developers*
Improper procedure to report issues with 24x7 system after hours
Long response from ServiceDesk
Routing incident ticket to FEF instead of SSA
Complex distributed system
Takes long time to realize what is wrong
Not streamlined rarely exercised stop/start procedure
dcache-core and dcache-pool enabled in runlevel (3)
Misconfiguration on installation
Improper start/stop procedure
Environment
Processes

Figure 2: Fishbone (Ishikawa) diagram that defines the problem and main cause categories with reasons that have contributed to the problem in each category.

Documentation on how to bring down/up dCache services are incomplete at CDF. This is a consequence of insufficient communication with dCache developers, staff rotation in CDF and in dCache developers group and recently changed support model.

`dcache-core` and `dcache-pool` should not be enabled in any runlevels to prevent system from starting prematurely. The fact that services were on in runlevel 3 indicate incomplete dCache installation by dCache developers.

In addition a lack of active monitoring that would clearly alert SSA primary what part of the system is not working. Currently it is mostly passive monitoring – a SAM shifter watches the monitoring plots or tables and alerts SSA if something does not look right.

Environment:

And finally it should be noted that the incident developed during off-hours and there was no SAM shifter on duty to watch monitoring information since the system was scheduled to be down during business hours.

System is distributed across multiple nodes. Without streamlines and well tested stop/start procedure it takes a long time to perform these operations and subsequently it takes a long tome to realize what is wrong with the system.

Root Cause

Incomplete dCache stop/start procedure and misconfiguration of local start up in runlevels resulted in non-functioning dCache system which was difficult do diagnose by CDF DH alone.

Contributing factors

There are several other factors contributed to the prolonged CDF dCache outage:

Technical:

1. PnfsManager does not reconnect indefinitely to dCache on connection failure, a feature of 1.7.0 and earlier releases of dCache.
2. KDC multiplexers did not start.

Procedural:

1. Improper stop/start procedure which was used due to :
    1. Lack of sufficient knowledge transfer from dCache developer liaison to CDF DH following modification of CDF dCache support model.
    2. Absence of documentation on how to properly stop/start CDF dCache
2. Misconfiguration of local dCache start up scripts setup overlooked by dCache installation procedure which resulted in dCache coming up in inconsistent state leading to flurry of erorrs that CDF DH expert could not interpret w/o external help.
3. Improper procedure of reporting incidents on 24x7 system to Service Desk during off hours.
4. Failure of Service Desk to direct incident ticket to SSA group promptly.
5. Lack of sufficient prioritization of incoming mails to SSA primary based on urgency.
6. Insufficient communication between CDF DH and dCache developers group about the scope and duration of CDF dCache downtime.
7. Lack of active monitoring of dCache components.

Recommendations

1. Service Desk has to rout incident tickets related to dCache to SSA group (already implemented)
2. Handling of high priority tickets by Service Desk needs to be reviewed and understood by all participants who are authorized to re-assign the tickets. Further delay with handling of ticket INC000000011514 might have been avoided if FEF representative either re-assigned the ticket to SSA or to Service Desk instead just replying to it with comment that it needs ot be reassigned to SSA.

3. CDF DH needs to use proper procedures established for 24x7 services to report dCache incidents during off hours..
4. CDF DH and dCache developers need to sit down and walk through existing at CDF documentation and scripts on how to start/stop and diagnose dCache. Update documentation and modify scripts as necessary.
5. Mails to dcache-admin@fnal.gov regarding dCache systems must be read and acted upon if necessary by dCache developer primary even if no bugzilla ticket is cut. Same is true for handling direct phone calls from CDF DH group.
6. SSA primary must handle Service Desk incident tickets with high priority.
7. SSA group must change default urgency setting for bugzilla tickets based on Service Desk incident tickets.
8. CDF DH must inform dCache developers and SSA group (dcache-admin@fnal.gov list) anytime a major disruption or change of configuration is planned to the system, so that a developer will be standing by to help in contingency situations.
9. Any prolonged downtime on dCache system must involve alerts of SSA by CDF DH (by sending e-mail to dcache-admin@fnal.gov) that they plan to shutdown dCache. CDF DH explicitly asks SSA to bring down PnfsManager. After shutdown is over CDF DH explicitly asks SSA to bring up PnfsManager. CDF DH checks http://cdfdca.fnal.gov:2288/cellInfo and alerts SSA if it stays OFFLINE within minutes after restart. At that point SSA alerts dCache developers.
10. Dcache developers must improve situation with active dCache monitoring to facilitate and expedite system diagnostics. Particularly the PnfsManager component must be monitored automatically and generate alerts (e.g. by sending mail to dcache-auto@fnal.gov) if this component is offline or non-responsive.
11. Dcache developers should disable dcache-core/dcache-pool in any run levels on CDF dCache nodes.
12. Dcache developers should modify kdcmux-boot to be compliant with chkconfig and enable it in run level 3 so that it boots up automatically in startup. Modify `cold-start/cold-stop` accordingly.

Root Cause Analysis Committee

The following people served on the RCA committee and/or contributed to the analysis of the incident:

Angela Bellavance
John Hendry
Stephan Lammel
Dmitry Litvintsev (chair)
Gene Oleynik
Timur Perelmutov
Carolina Sinclair
Rick Snider
Margaret Votava
Eric Wicklund